NISTIR 8324

2018 Media Forensics Challenges (MFC18): Summary and Results

Yooyoung Lee
Amy N. Yates
Haiying Guan
Andrew Delgado
Daniel Zhou
Timothee Kheyrkhah
Jonathan Fiscus

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8324

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8324

2018 Media Forensics Challenges (MFC18): Summary and Results

Yooyoung Lee
Amy N. Yates
Haiying Guan
Andrew Delgado
*Daniel Zhou
Timothee Kheyrkhah
Jonathan Fiscus
Information Technology Laboratory
Information Access Division

**Former employee; all work for this publication was performed while working at NIST*

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8324

November, 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology
Interagency or Internal Report 8324
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8324, 27 pages (November, 2020)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8324

Abstract

The interest in forensic techniques capable of detecting many different media manipulation types has been growing, and system development with machine learning technology has been evolving in recent years. There has been, however, a lack of diversity in the data collections and in the evaluation methodologies for advancing multimedia forensics technologies. For the forensics research community, a well-defined evaluation is necessary to rapidly measure the accuracy and robustness of systems over diverse datasets collected under various environments. In this paper, we propose an evaluation framework and associated performance metrics and apply them to the 2018 Multimedia Forensics Challenge (MFC18). This MFC18 evaluation consists of five tasks and two challenges. A large number of datasets were created to support each task and for conducting the experiments using a structured evaluation framework. A total of 25 teams participated in the MFC18 evaluation; we analyse their performance on the tasks and challenges, and provide performance rankings for each team's best-performing system.

Key words

forensics, evaluation, media authenticity, tamper detection, accuracy, and robustness.

Table of Contents

1 Introduction **1**

2 Evaluation Tasks and Measures **3**
- 2.1 Task Definition 3
- 2.2 Performance Measures 4

3 Evaluation Framework **6**
- 3.1 Evaluation Types 6
- 3.2 Evaluation Conditions 6
- 3.3 Scoring Framework 7

4 Dataset **8**
- 4.1 Journal Type 8
- 4.2 Data Distribution 11

5 Results **14**
- 5.1 Full Scoring Framework 14
- 5.2 Selective Scoring Framework (MDL-Image only) 21
- 5.3 NC17 and MFC18 Comparison 22

6 Conclusions **24**

References **25**

List of Tables

Table 3.1 A Summary of the MFC18 Scoring Frameworks 7

Table 4.1 Summary of the MFC18 Evaluation Tasks Datasets[a] 12

Table 4.2 MFC18 Camera Verification (CV) Challenge Dataset 13

Table 4.3 MFC18 Generative Adversarial Networks Detection (GAN) Challenge Dataset 13

Table 5.1 Summary of MFC18 Image/Video Five Evaluation Task Results (Full Scoring Only) 16

Table 5.2 Summary of MFC18 Image/Video Challenge Results (Full Scoring Only) 17

List of Figures

Figure 1.1 MFC18 dataset examples for image and video; (a) genuine image; (b) image manipulation is done by removing, splicing, cloning, among others; (c) genuine video; (d) video manipulation is done by replicating with CGI technique. 1

Figure 2.1 Performance measures (a) Receiver Operating Characteristic (ROC) and Area Under Curves (AUC) for detection and verification metrics; (b) Computation of mask confusion matrix for localization metrics [1]. 5

Figure 4.1 Examples of the three journal types; (a) Human-JT denotes the set of journals created by people manually manipulating media using editing tools; (b) Extended-JT is the set of journals created by an algorithm taking existing journals and extending them by automatically creating new manipulated media; (c) Auto-JT is the set of journals automatically generated by an Auto-JT tool to produce a large corpus of manipulated images by a specified design. 9
Figure 4.2 Distribution of the three journal types (MDL case); the stack histogram shows the Dev1 and Dev2 training sets have unbalanced distribution across three journal types while the EvalPart1 test set is balanced with Human-JT (30%), Extended-JT (31%), and Auto-JT (39%), respectively. 10
Figure 4.3 MFC18 Manipulation types distribution (MDL-Image); The *x*-axis denotes the manipulation types, and the *y*-axis is the stacked instance counts of Dev1 (blue), Dev2 (orange), and EvalPart1 (green) datasets for each manipulation type. Note the unbalanced manipulation type distributions across the three datasets. 11
Figure 4.4 Manipulation type distribution on GAN image dataset; both GAN-based (marked in red) and non-GAN manipulation types were used for the target trial and unbalanced manipulation types was used in the MFC18 evaluation. 13
Figure 4.5 An example of the MFC18 GAN dataset (approved by IRB ITL-0018); (a) genuine image; (b) the manipulated image contains both GAN (e.g., ErasureByGAN) and non-GAN (e.g., ColorBalance) operations, so the authors suggest interpreting the GAN challenge results with caution. 14
Figure 5.1 An example of ROC curves for participants in the MFC18 MDL-Image detection task; the x-axis is a false alarm rate (FAR) and the y-axis is a correct detection rate (CDR). 15
Figure 5.2 A graphical example of localization evaluations; top-left: ground-truth masks marked in color by different manipulation types, top-right: system output mask, bottom-left: Mask confusion matrix visualization, bottom-right: confusion matrix results for scoring localization. 15
Figure 5.3 System performance ranking for MDL-Image; (a) detection performance ranking ordered by AUC; (b) localization performance ranking (OptMCC indicates MCC_o); the highest-performed system in detection is not necessary the highest-performed system in localization. 17
Figure 5.4 Effect of journal types across systems (MDL-Image case); The *x*-axis is the team and the *y*-axis is the mean AUC. The characters inside each bar represent the settings of the journal types. The journal type has a larger effect on some systems (e.g. T10, T18, T24) but is not as noticeable for other teams (e.g., T07 and T20). 18

Figure 5.5 System performance ranking for MDL-Video; (a) detection performance ranking ordered by AUC; (b) localization performance ranking ordered by MCC_o; The team T07 has the highest AUC (0.59) and the highest MCC (0.09). 19
Figure 5.6 System performance ranking for SDL; (a) detection performance ranking ordered by AUC; (b) localization performance ranking ordered by $pMCC_o$ (ProbeOptMCC and DonorOptMCC indicate $pMCC_o$ and $dMCC_o$, respectively.); The team T04 has the highest performance for both detection ($AUC = 0.77$) and localization ($pMCC_o = 0.36$, $dMCC_o = 0.33$). 19
Figure 5.7 The EV data and system performance ranking; (a) data distribution for MFC18 EvalPart1 and training data; (b) verification performance ranking ordered by AUC; The team T03 has the highest AUC value (0.85). 20
Figure 5.8 System performance ranking for provenance tasks; (a) PF performance ranking by $Recall@300$; (b) PGB performance ranking ordered by sim_{NLO}; The team T09 has the highest performance for both PF ($Recall@300 = 0.90$) and PGB ($sim_{NLO} = 0.57$). 20
Figure 5.9 CV System performance ranking for CV; (a) Image-Image performance ranking ordered by AUC; (b) Video-Video performance ranking ordered by AUC; T15 has the highest ($AUC = 0.87$) for Image-Image while T20 has the highest ($AUC = 0.70$) for Video-Video. 21
Figure 5.10 System performance ranking for GAN; (a) GAN-Image performance ranking ordered by AUC; (b) GAN-Video performance ranking ordered by AUC; note that probes contain multiple manipulation types other than GAN-related operations which require caution to interpret the GAN challenge results. 22
Figure 5.11 System performance ranking for selective scoring (MDL-Image: Clone); (a) detection performance ranking ordered by AUC; (b) localization performance ranking; T24 has the highest performance for both detection ($AUC = 0.81$) and localization ($MCC_o = 0.27$). 23
Figure 5.12 System performance ranking for selective scoring (Remove and Crop); (a) Remove detection performance ranking ordered by AUC; (b) Crop detection performance ranking ordered by AUC; T07 has the highest performance for both the Remove and Crop detection ($AUC = 0.81$ and $AUC = 0.83$, respectively). 23
Figure 5.13 Tasks performance comparison of the NC17 and MFC18 evaluations (different systems and different datasets were used for the comparison). 24

Glossary

Acronyms

CV Camera Verification. 2–4, 12–14, 16, 17, 20, 21, 24

EV Event Verification. 2–4, 11, 12, 14, 16, 19, 20, 24

GAN Generative Adversarial Networks Detection. 2–4, 12, 13, 16, 17, 21, 22, 24

MDL Manipulation Detection and Localization. 2–4, 11, 12, 14–19, 21–24

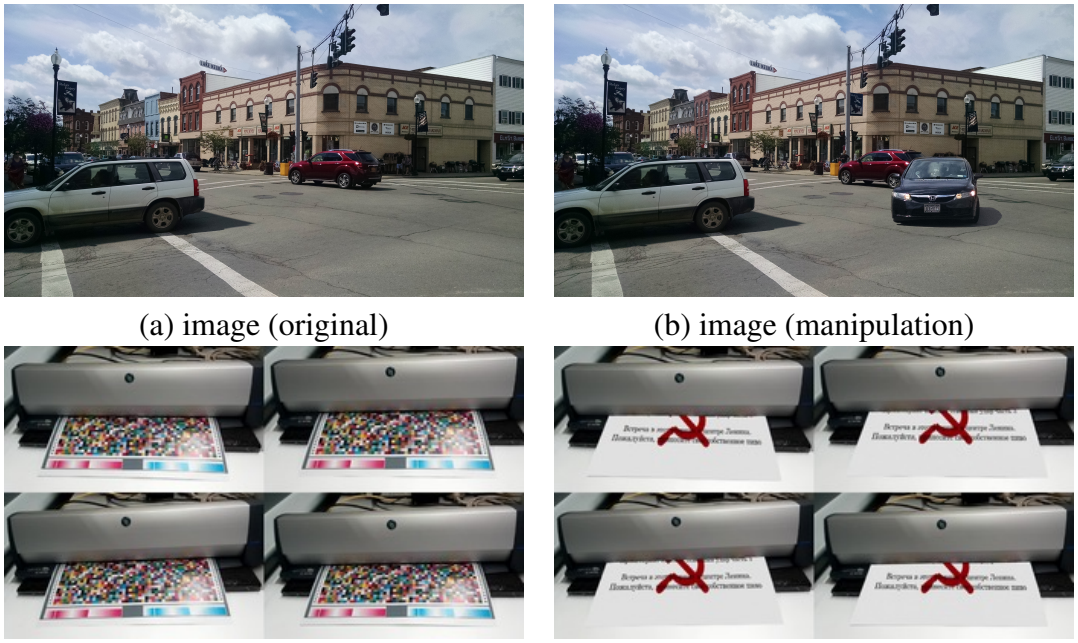
PF Provenance Filtering. 2–5, 12, 16, 20, 22–24

PGB Provenance Graph Building. 2–5, 12, 16, 20, 24

SDL Splice Detection and Localization. 2–4, 11, 12, 14, 16, 18, 19, 22, 24

1. Introduction

Recent advancements have produced a large variety of media editing tools that are easily accessible. The result is that digital media content (e.g., image, video, and audio) can be easily altered, falsified, and redistributed, sometimes for legitimate purposes [2, 3]. The volume of multimedia is large, however, the fidelity of the media is no longer by default trustworthy, including media obtained from sources such as online news, magazines, social media, and even prestigious journals [4]. There is growing interest in universal forensic techniques capable of detecting many different editing operations [5, 6]. Developments of these systems along with machine learning techniques have been evolving in recent years. The evaluation of such systems is vital to advance the forensic technologies. However, previous studies have been limited to a single manipulation evaluation to measure the accuracy and robustness of the systems in the forensic applications. In spite of the wide variety of possible manipulation operations on the tampered media, the evaluations were often focused on a single processing manipulation [5, 7, 8].



(a) image (original) (b) image (manipulation)

(c) video (original) (d) video (manipulation)

Figure 1.1. MFC18 dataset examples for image and video; (a) genuine image; (b) image manipulation is done by removing, splicing, cloning, among others; (c) genuine video; (d) video manipulation is done by replicating with CGI technique.

Most widespread manipulation methods fall into a small number of categories, including adding (e.g. splicing), replicating (e.g. cloning), or removing (e.g. seam carving) content. Such manipulations [1] can be performed in easily accessible editors (e.g., Adobe Photoshop, GIMP, or ImageMagick), resulting in output images that are very realistic. Detection of such high quality manipulations can be challenging even for forensic analysts.

The advent of autoencoders (AE) and generative adversarial networks (GAN) enabled the creation of fake media with unprecedented levels of realism (e.g., deep fakes) [9, 10], which drew attention in the media forensics community.

In response to these developments, in 2017, the Defense Advanced Research Projects Agency (DARPA) initiated a series of media forensics (MediFor) evaluations, conducted by the National Institute of Standards and Technology (NIST). The primary goal of the evaluations is to advance media forensics technologies that can automatically determine media authenticity as well as the history of the images that have been manipulated and its relationship.

This paper is a summary of 2018 Media Forensics Challenge (MFC18) results and how evaluation and analysis methodologies are applied to MFC18. Figure 1.1 illustrates original images/videos as well as the manipulated images/videos appeared in the MFC18 dataset.

MFC18 develops a task-driven evaluation approach and consists of five tasks and two challenges. The tasks are Manipulation Detection and Localization (MDL), Splice Detection and Localization (SDL), Event Verification (EV), Provenance Filtering (PF), and Provenance Graph Building (PGB). The challenges are Camera Verification (CV), and Generative Adversarial Networks Detection (GAN).

This paper is a summary of MFC18 results and how evaluation and analysis methodologies are applied to MFC18. There are several key impacts and contributions from MFC18; specifically, it:

- develops an evaluation structure using both a full scoring framework and a selective scoring framework for accessing a wide domain of manipulation operations,
- defines the evaluation tasks and methods as well as their performance metrics for targeting media forensics applications,
- produces a large number of MFC18 datasets for supporting both a general evaluation and a task-driven evaluation,
- conducts comparative analysis using a structured evaluation protocol for accessing an accuracy and robustness of a system,
- initiates the first GAN challenge in media forensics evaluation community, and
- provides constructive design and analysis recommendations for improvements of future media forensic evaluations.

The paper is organized as follows: the next section defines the MFC18 evaluation tasks and challenges and describes the proposed performance measures. The evaluation methods and the datasets used are described in Secs. 3 and 4, respectively. Finally, in Secs. 5 and 6, we provide results for each task and conclude the paper with the summary results and the findings.

2. Evaluation Tasks and Measures

This section provides a brief overview of MFC18 evaluation tasks/challenges and evaluation metrics. The MFC18 evaluations were primarily performed on five tasks (MDL, SDL, EV, PF, and PGB) and two challenges (CV, GAN). In this paper, a *"base"* indicates original media with high provenance while a *"probe"* indicates manipulated media. A *"donor"* indicates another media source whose region(s) were spliced into the probe. The MFC18 evaluation plan [1] includes a detailed description of the tasks/challenges and performance metrics.

2.1 Task Definition

The *Manipulation Detection and Localization (MDL)* task was to detect if the media (i.e., image or video) had been manipulated and, if so, then to spatially (temporally for video) localize the manipulated region. For detection, an MDL system provided a confidence score for each trial with higher numbers indicating the media was more likely to have been manipulated. For the localization evaluation, the system provided a mask and its bit plane that indicated the manipulated region(s) with a manipulation type. Local manipulations (e.g., clone) required a mask output, while global manipulations (e.g., blur) affecting the entire media did not require a mask. The MDL task was divided into two probe sets: an image probe set and a video probe set.

The *Splice Detection and Localization (SDL)* task was to detect if a region of a given image (i.e., the donor) had been spliced into another image (i.e., the probe) and, if so, then to localize the region(s) of the donor and probe images that were used for the splice operation. Similar to the MDL task, a SDL system provided a confidence score along with two masks: one with the region(s) of the donor that was copied and another with the region(s) of the probe that was pasted from the donor.

The *Event Verification (EV)* task was to determine if an image is associated with a claimed event, given a collection of images and videos from the event. A EV system provided a confidence score; image localization was not evaluated for this task.

The *Provenance Filtering (PF)* task was defined as searching for a potential pool of related images (that may be present in a phylogeny graph [11, 12] with respect to the given probe image) from a large collection of images (called the world dataset). Given a probe image, the goal of the PF task was to return up to 500 images of the predicted ancestors and descendants including the original base images from the provided world dataset. A PF system provided a JSON file that contained 500 filtered images (represented as nodes) including the given probe image itself and a confidence score for each node that indicated how likely the filtered image was related with respect to the probe image.

The *Provenance Graph Building (PGB)* task was retrieving images related to the given probe image from the world dataset and constructing the relationships among the retrieved images. This includes finding the ancestor and descendent sequences. The probe image could be a base, donor, intermediate, or final modified image. The goal of this task was to construct a provenance phylogeny graph (for the given probe image) that described the relationships among the associated images with the manipulation sequences. A PGB sys-

tem provided a JSON file that contained both nodes and links with the two types of confidence scores: (1) a confidence score for each node indicating how likely the retrieved image (node) was present in the provenance graph of the probe image and (2) a confidence score for each link indicating how likely the two nodes between a source node and a target node have the relationship (link) in the provenance graph. Although PGB system was required to provide a full provenance graph, for MFC18, we evaluated the system under two conditions: (1) full-set graph and (2) subset graph. For the full-set graph condition, all images related to the probe image were evaluated with the ancestors and descendants' sequences, while for the subset graph condition, the node set (the subset of the related images) was restricted to ancestors and descendants of the probe image and only directed paths related to the probe image were evaluated. In this paper, we present the results with the full-set graph condition only.

The *Camera Verification (CV)* challenge was to determine if a camera fingerprint from an image matches a claimed camera fingerprint, given a collection of camera device IDs. This task supported both image and video probes, and the dataset consisted of three training sets (image, video, and multimedia) and two testing sets (image and video). This yielded a total of six training-testing conditions with the composition of the training sets and the testing sets (see Table 4.2). A CV system provided a confidence score indicating how likely the image was captured with the claimed camera, and the system was required to indicate the training-testing condition for the submission.

The *Generative Adversarial Networks Detection (GAN)* challenge was to evaluate if a system detected manipulated images/videos created by a generative model (called GAN-based manipulations). With recent advances in GAN (Generative Adversarial Network) techniques [13], realistic fake objects (e.g., faces) on media can be generated. To the best of our knowledge, MFC18 was the first evaluation to address such a GAN challenge [9]. For the GAN challenge, the system was allowed to adapt the generative model for their system development—the nature of the system for the GAN challenge may have been different with the MDL task. To compare system performance between the GAN-based manipulations and the other manipulations (non-GAN), both the GAN and non-GAN manipulations are evaluated in the challenge.

2.2 Performance Measures

For *detection and verification* evaluations, system performance was measured by Area Under Curve (AUC) and Correct Detection Rate at a False Alarm Rate of 5% (called CDR) from the Receiver Operating Characteristic (ROC) [14] as shown in Figure 2.1-a.

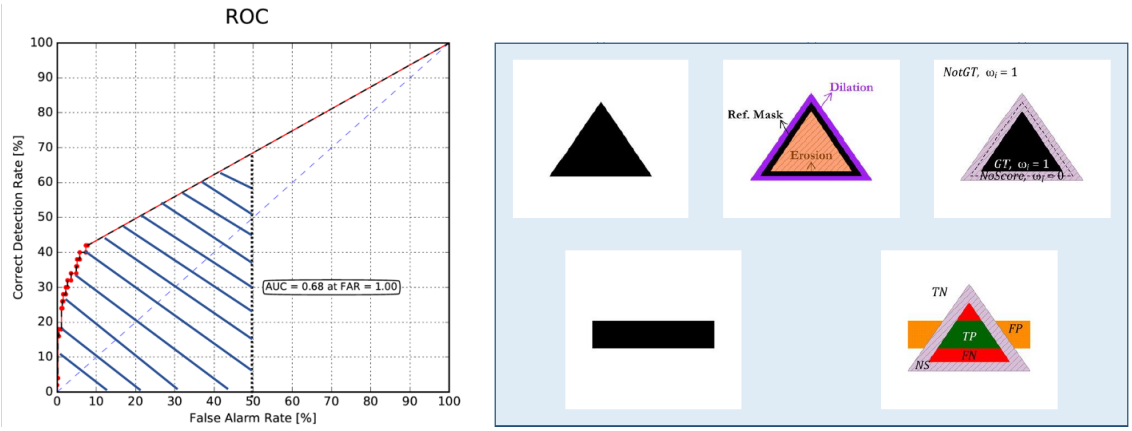
For *localization* evaluations, the Matthews Correlation Coefficient (MCC) [15] was primarily used. The optimum MCC (MCC_o) was calculated using an ideal mask-specific threshold found by computing metric scores over all pixel thresholds t. Figure 2.1-b shows a visualization of the different mask regions used for image mask evaluations.

- $TP(t)$ (True Positive) is the overlap area of the reference mask and system output mask labelled as manipulated at the pixel threshold t(green).

- $FN(t)$ (False Negative) is where the reference mask indicates the area as manipulated, but the system did not detect it as manipulated at the threshold t (red).
- $FP(t)$ (False Positive) is where the reference mask indicates the region is not-manipulated, but the system detected it as manipulated at the threshold t (orange).
- $TN(t)$ is the where reference mask indicates not-manipulated, and the system also detected it as not-manipulated at the threshold t (white).
- NS (No-Score) is the region of the reference mask not scored (purple), the result of the dilation and erosion operations.

If the denominator is zero, then $MCC_o = 0$. If $MCC_o = 1$, there is perfect correlation between the reference and system output masks. If $MCC_o = 0$, there is no correlation between the reference and system output masks. If $MCC_o = -1$, there is perfect anti-correlation.

For videos, manipulation detection and temporal localization (not including spatial localization) were evaluated in MFC18. The objective was to detect if a video has been manipulated and then to determine which time segments have been manipulated. *MCC* was calculated on the time segments.



(a) Detection metrics (b) Localization metrics

Figure 2.1. Performance measures (a) Receiver Operating Characteristic (ROC) and Area Under Curves (AUC) for detection and verification metrics; (b) Computation of mask confusion matrix for localization metrics [1].

For the PF task, recall of the top-k retrieved images was primarily used to examine system performance. For the MFC18 evaluation, we used the list of $k \in \{50, 100, 200, 300\}$ which are represented as recall@50, recall@100, recall@200, and recall@300, respectively. For the PGB task, we adopted the graph similarity metrics proposed by Papadimitriou *et al.* [16]. The performance metrics determine the overlap of nodes (vertices) and

links (edges) between the reference and the system output provenance graph and are defined as follows:

$$sim_{NO} = 2\frac{|V_r \cap V_s|}{|V_r| + |V_s|} \tag{1}$$

$$sim_{LO} = 2\frac{|E_r \cap E_s|}{|E_r| + |E_s|} \tag{2}$$

$$sim_{NLO} = 2\frac{|V_r \cap V_s| + |E_r \cap E_s|}{|V_r| + |V_s| + |E_r| + |E_s|} \tag{3}$$

where $G_r = (V_r, E_r)$ is the reference provenance graph, V_r is the set of nodes, and E_r is the set of links for the reference while $G_s = (V_s, E_s)$ is the system provenance graph, V_s is the set of nodes, and E_s is the set of links for the system output. sim_{NO} is the overlap of nodes, sim_{LO} is the overlap of links, and sim_{NLO} is the overlap of both nodes and links. If $G_s = G_r$, then $sim_{NO} = sim_{LO} = sim_{NLO} = 1$.

For all the metrics described above, a higher value is considered as better performance.

3. Evaluation Framework

This section gives an overview of the evaluation type, condition, and scoring framework that were used in the MFC18 evaluation.

3.1 Evaluation Types

For the MFC18 evaluation, there were the two evaluation types: 1) open evaluation and 2) sequestered evaluation. For the open evaluation, the performers ran their software on their hardware and configurations, and submitted the system output with the defined format. On the other hand, for the sequestered evaluation, the performers submitted their run-able system, and system performance was independently evaluated on the sequestered data using the evaluator's hardware. The following report and analysis are based on the results from the *open evaluation* only.

3.2 Evaluation Conditions

In the MFC18 evaluation, systems provide their outputs with an option of four different conditions: 1) image-only, 2) image-and-metadata, 3) video-only, and 4) video-and-metadata. For the image-only condition, the system was only allowed to use the pixel-based content of images as input. On the other hand, for the image-and-metadata condition, the system was allowed to use metadata, including image header, in addition to the pixel-based content for the image. For the video-only condition, similar to the image conditions, the system was only allowed to use the pixel-based content for videos and audio if it existed as input, while for the video-and-metadata condition, the system was allowed to use metadata,

including video header or other information. In this paper, we primarily focus on condition 1 and 3; *image-only* and *video-only*, respectively.

3.3 Scoring Framework

With advanced editing and deep learning tools, realistic tampered images/videos with diverse manipulation techniques (e.g., remove, clone, splice, crop, synthesized media, among others) are widely available everywhere. Although the goal of the MFC18 evaluation is to develop a general forensic application, it is often a challenge for a system to attack all of the diverse manipulation operations. Actual cases that forensic analysts face every day are different and situational, and sometimes forensic analysts need a robust tool that can support a specific application domain by taking advantage of a suite of tampering techniques proven to be effective in that domain. For example, system targeting face-swap detection may not be able to process a tampered image where a face is not present.

In this regard, the MFC18 evaluation supported three different scoring frameworks: (1) full scoring, (2) opt-in scoring, and (3) selective scoring. Table 3.1 is a summary of the scoring frameworks.

Table 3.1. A Summary of the MFC18 Scoring Frameworks

Det: Detection, Loc: Localization, Target: Target manipulation trials to be detected, Non-Target: Original media or non-target manipulation trials

	Full Scoring		**Opt-In Scoring**		**Selective Scoring**	
	Det	Loc	Det	Loc	Det	Loc
Target	Full Set	Full Set	Subset (Status)	Subset (Status)	Subset (Query)	Subset (Query)
Non-Target	Full Set		Subset (Status)		Full Set	

For *full scoring*, performance was scored over a full test set from all trials (for both a target and non-target set) present in the task, regardless of the trial's process status or the intended manipulation type detection specified by the system. For localization, performance was scored on detected target trials only.

Another scoring framework was *opt-in scoring*. For all tasks, a system could use a strategy to identify which probes were appropriate for a response. The system indicated for each trial if it was appropriate for the system to respond for the probe, if it was not appropriate for the system to respond, or if it was only appropriate for the system to respond to one sub-task (e.g., detection but not localization). For localization sub-tasks, specific portions of an image mask were also able to be deemed not appropriate for a response, denoting the region with specific pixel value. Even when a system determined it was not appropriate to respond, the system was required to enter a response, which was scored under full scoring. On the other hand, under opt-in scoring, only trials the system deemed appropriate for a response on the task or sub-task were scored. Both target and non-target trials were subsetted, based on the system's determination. Under opt-in scoring, in addition to performance

metrics, the fraction of processed trials executed by the system was reported, known as the trial response rate (TRR).

The third scoring framework was *selective scoring*, which used a subset (query) that queried manipulation types from the target trials only. It could be used to filter the target trials for manipulations of interest, and it was scored on the subset of target trials while using all non-target trials (fixed number of non-targets). The selective scoring protocol could also be utilized for localization; mask regions containing non-selected manipulation types were treated as no-score regions for measuring localization performance; both full scoring and the opt-in scoring were available for the selective scoring results. Details of the evaluation framework are available in the MFC18 evaluation plan [1].

For simplicity, in this paper we present the results for *full scoring* and three instances of *selective scoring* only.

4. Dataset

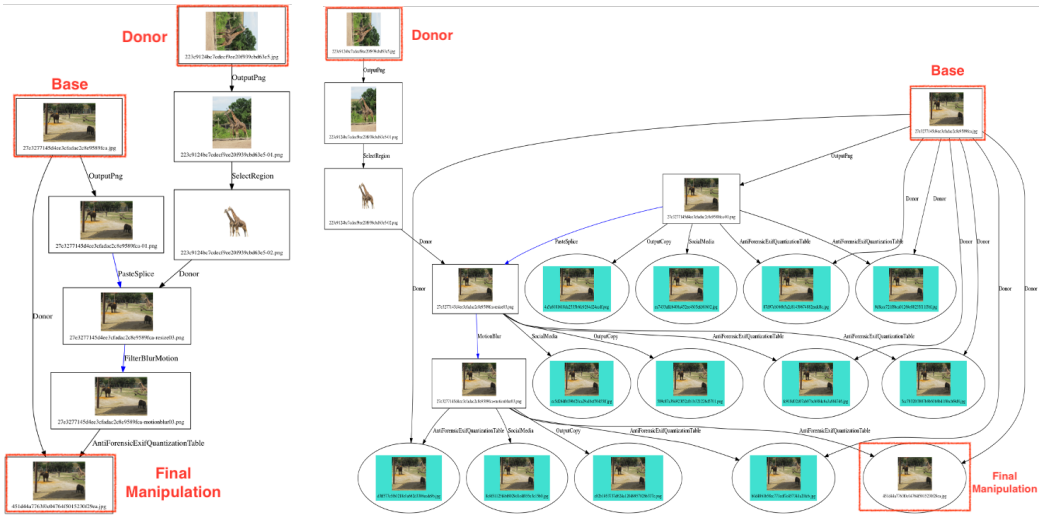
For the MFC18 data collection, an original image or video (called a base) was used to generate the diverse manipulated media collections. A donor media was defined as a media that had been spliced into a target manipulated media. While creating manipulations given a base image or video, it was important to keep track of each manipulation step and to generate a mask for that manipulation as a unit output (called a journal). Thus, PAR Government Systems in collaboration with NIST developed a Journaling Tool (JT) by Robertson et al. [17] that created a graph of manipulation history and its segmentation masks for the unit output of a graph, a base , donor media, and manipulated media) as shown Figure 4.1-a.

4.1 Journal Type

As illustrated in Figure 4.1, the MFC18 dataset consists of the three journal types for the data collections, namely: (a) Human-JT, (b) Extended-JT, and (c) Auto-JT. A journal type is a set of journals that all share a property.

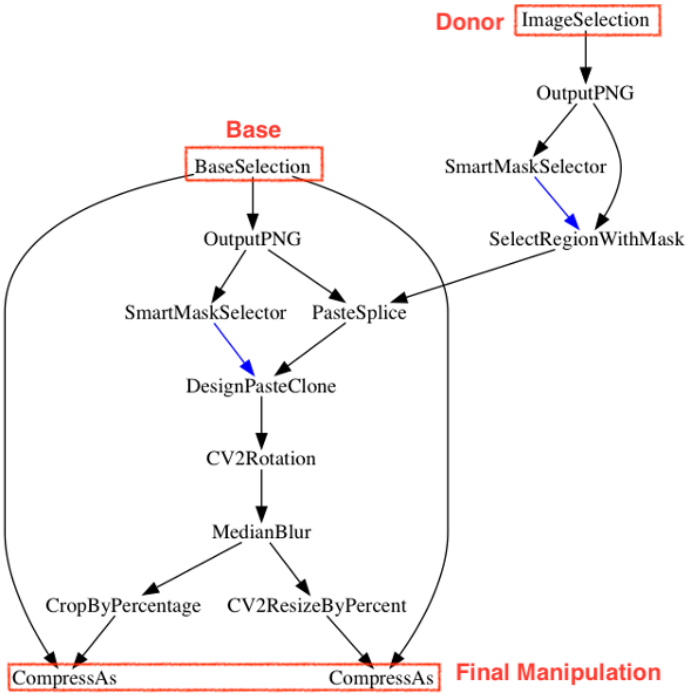
Human-JT denotes the set of journals created by people manually manipulating media using editing tools and software, including but not limited to Adobe Photoshop, ImageMagick, and GIMP. As illustrated in Figure 4.1-a, a node in the graph represents a piece of media (an image in this example) and a directed link indicates what manipulation type was applied to create the following target image from the previous image. Human-based manipulations are often expensive and time consuming, and the data sample size may not be statistically balanced and enough for evaluating the accuracy and robustness of a system. Thus, we provided the other tools that automatically created manipulations.

Extended-JT denotes the set of journals created by an Extended Journaling Tool taking existing journals and extending them by automatically creating new manipulated media based on the nodes in the existing journal. Extended-JT augments the collection of existing journals for supporting or expanding necessary target manipulations in the scope of research interests. The extended manipulations are marked in green on Figure 4.1-b.



(a) Human-JT

(b) Extended-JT



(c) Auto-JT (Random Design case)

Figure 4.1. Examples of the three journal types; (a) Human-JT denotes the set of journals created by people manually manipulating media using editing tools; (b) Extended-JT is the set of journals created by an algorithm taking existing journals and extending them by automatically creating new manipulated media; (c) Auto-JT is the set of journals automatically generated by an Auto-JT tool to produce a large corpus of manipulated images by a specified design.

One drawback of human journals is that there are limitations to cover diverse manipulations; the overall data size may be insufficient for train and evaluation as well as it contains the unbalanced manipulation types. To support such limitations, the Automated Journaling Tool was developed to support a large number of datasets (that mimic human-based journals) and to support a wide domain of manipulation types (balanced sample size). The tool is capable of creating diverse manipulations by applying a manipulation layer design. For the MFC18 evaluation, we used three different kinds of layered design: (1) single layer, (2) k layers, and (3) random number of layers. Auto-JT denotes the set of journals created by the fully-automated tool without human intervention. Figure 4.1-c is an Auto-JT journal type example from the auto-generated random layer design in the MFC18 data.

Using the three journal types, a large scale of the datasets (for both image and video) was created to support the MFC18 evaluations.

Figure 4.2 is a summary distribution of the three journal types for image datasets. The EvalPart1 dataset is a test set for evaluation and the Dev1 and Dev2 indicates a training set for supporting system development. We included a large number of the Auto-JT journals for the Dev2 training set. The training sets (Dev1 and Dev2) have unbalanced distribution across the three journal types. The proportion among the three journal types is, however, balanced for the EvalPart1 test set: Human-JT (30%), Extended-JT (31%), and Auto-JT (39%), respectively.

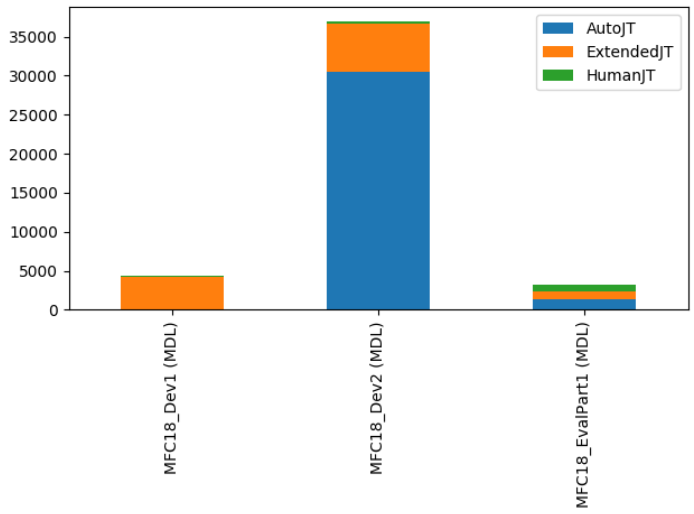


Figure 4.2. Distribution of the three journal types (MDL case); the stack histogram shows the Dev1 and Dev2 training sets have unbalanced distribution across three journal types while the EvalPart1 test set is balanced with Human-JT (30%), Extended-JT (31%), and Auto-JT (39%), respectively.

4.2 Data Distribution

Table 4.1 is a summary of the datasets for each task. Again, the Dev1 and Dev2 are indicated as a training set (incrementally released as the subsets of the training sets). The test set was divided into the three subsets: EvalPart1, EvalPart2, and EvalPart3. The EvalPart1 dataset was used for open evaluation while the EvalPart2 and EvalPart3 datasets were used for sequestered evaluation. The paper written by Guan et al. [18] describes the MFC18 dataset collections in details. Since we address the *open evaluation only* in this paper, we present our results based on the *EvalPart1 dataset only*.

Figure 4.3 shows the distribution of the manipulation types that were used for the MDL-Image task. The *x*-axis denotes the manipulation types, and the *y*-axis is the stacked instance counts of Dev1 (blue), Dev2 (orange), and EvalPart1 (green) datasets for each manipulation type. The distribution of the manipulation types are unbalanced.

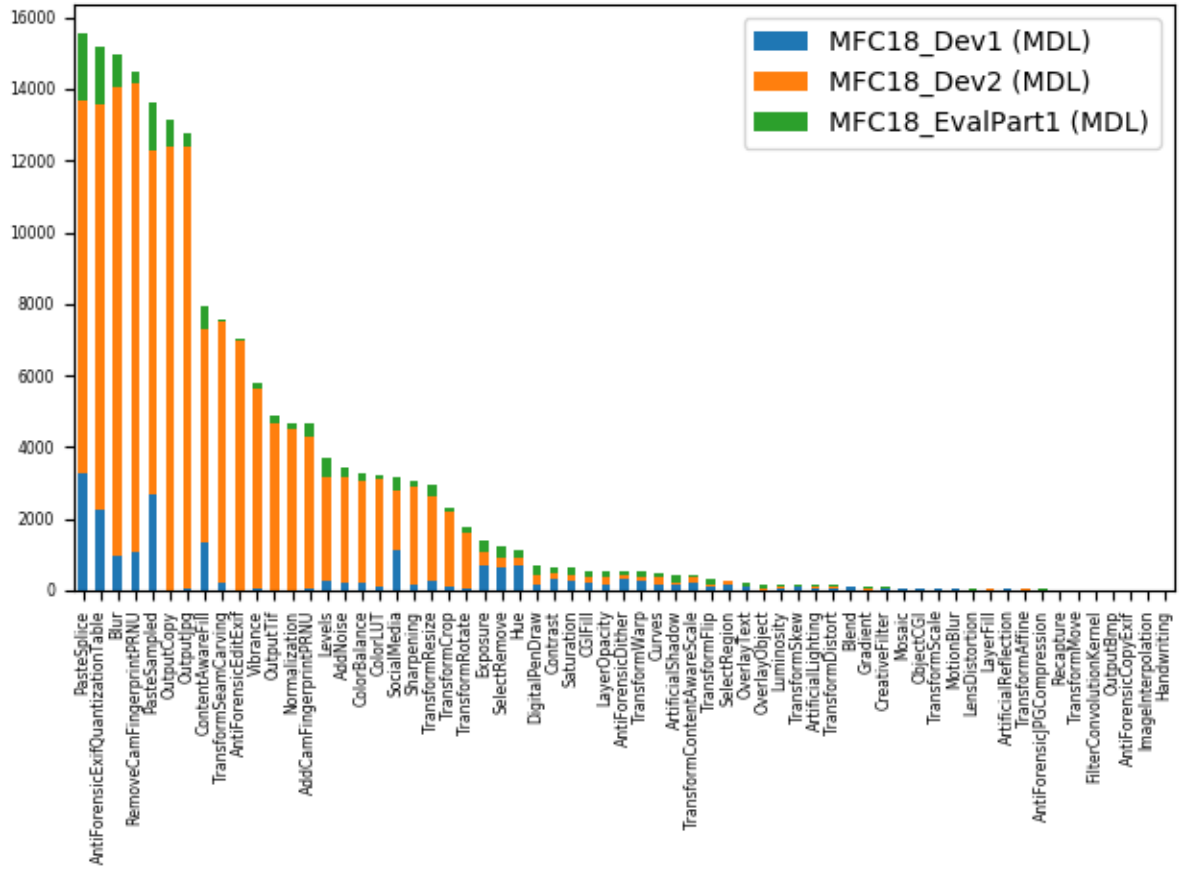


Figure 4.3. MFC18 Manipulation types distribution (MDL-Image); The *x*-axis denotes the manipulation types, and the *y*-axis is the stacked instance counts of Dev1 (blue), Dev2 (orange), and EvalPart1 (green) datasets for each manipulation type. Note the unbalanced manipulation type distributions across the three datasets.

The Event Verification (EV) task is a verification task while MDL and SDL are detec-

tion tasks. The target and non-target trials are therefore a pair-event, and the #N and #NT in Table 4.1 (EV column) are the number of the pair-event trials.

For the Camera Verification (CV) task, the MFC18 dataset was divided into the six subsets depending on the training-testing conditions. Table 4.2 shows the training-testing conditions and their relevant numbers of target and non-target trials. In this paper, for simplicity, we primarily discuss the results when the training and testing sets are on the same media type; namely, Image-Image and Video-Video.

As mentioned in Sec. 2.2, the Generative Adversarial Networks Detection (GAN) challenge allowed the teams to adapt the generative model for their system development. As illustrated in Table 4.3, we included a GAN image dataset with 855 target trials (mixed of GAN and non-GAN manipulations) and 485 non-target trials, and a GAN video dataset with 50 target trials and 68 non-target trials to compare system performance.

For the GAN image dataset, the GAN-based manipulations can be categorized into local-GAN and global-GAN. The local-GAN operations included the GANFill and GAN-based PasteSplice. For instance, the donor image generated by the GAN model pastes into the local area of the probe image. On the other hand, the global-GAN contains Erasure-ByGAN and anti-forensics (based on camera model anti-forensic techniques generated by GAN). The definition of manipulation operation types is described in the user guide written by E.Robertson [19]. Note that the crop and resize operations generated by either GAN or non-GAN tools are not a part of the target trials for the GAN challenge.

Figure 4.4 illustrates a distribution of the manipulation operations for the GAN image dataset. A major drawback in this GAN dataset is that GAN manipulations were combined with non-GAN manipulations in the probe set (as shown Figure 4.4) and design of the dataset required further examination. For instance, Figure 4.5 is a sample from the MFC18 GAN image dataset; this manipulated image contains both GAN (e.g., ErasureByGAN), and non-GAN (e.g., ColorBalance) operations, so the authors suggest interpreting the GAN challenge results (or even using the GAN challenge dataset) with caution.

Table 4.1. Summary of the MFC18 Evaluation Tasks Datasets[a]

Dev1 and Dev2 are the training set and EvalPart1 is the test set. #T indicates the number of the target trials and #NT is the number of the non-target trials. #Probe is a number of the images or videos that are the subject of the task question posed to the system. #World is the number of images and videos pool that simulates a real-world collection of unknown provenances.

MFC18 Datasets	Image						Provenance (PF + PGB)		Video	
	MDL		SDL		EV				MDL	
	#T	#NT	#T	#NT	#T	#NT	#Probe	#World	#T	#NT
Dev1	4,395	1,162	2,496	47,504	400	1,200	209	13,304		
Dev2	36,910	1,429	10,009	39,991	398	1,194	318	31,005		
EvalPart1	3,265	14,156	1,105	15,618	600	6,600	1,122	1,020,339	323	713

Table 4.2. MFC18 Camera Verification (CV) Challenge Dataset

The MFC18 dataset was divided into subsets depending on the training-testing conditions. The primary focus in this paper is the conditions from Image-Image and Video-Video only.

Condition		#T	#NT
Train	Test		
Image	Image	2,440	2,835
Video	Image	1,720	1,663
Multimedia	Image	1,720	1,663
Image	Video	101	188
Video	Video	101	188
Multimedia	Video	101	188

Table 4.3. MFC18 Generative Adversarial Networks Detection (GAN) Challenge Dataset

#T indicates the number of the target trials and #NT is the number of the non-target trials.

Image			Video	
#T		#NT	#T	#NT
Local/Global GAN	Non-GAN			
729	126	485	50	68

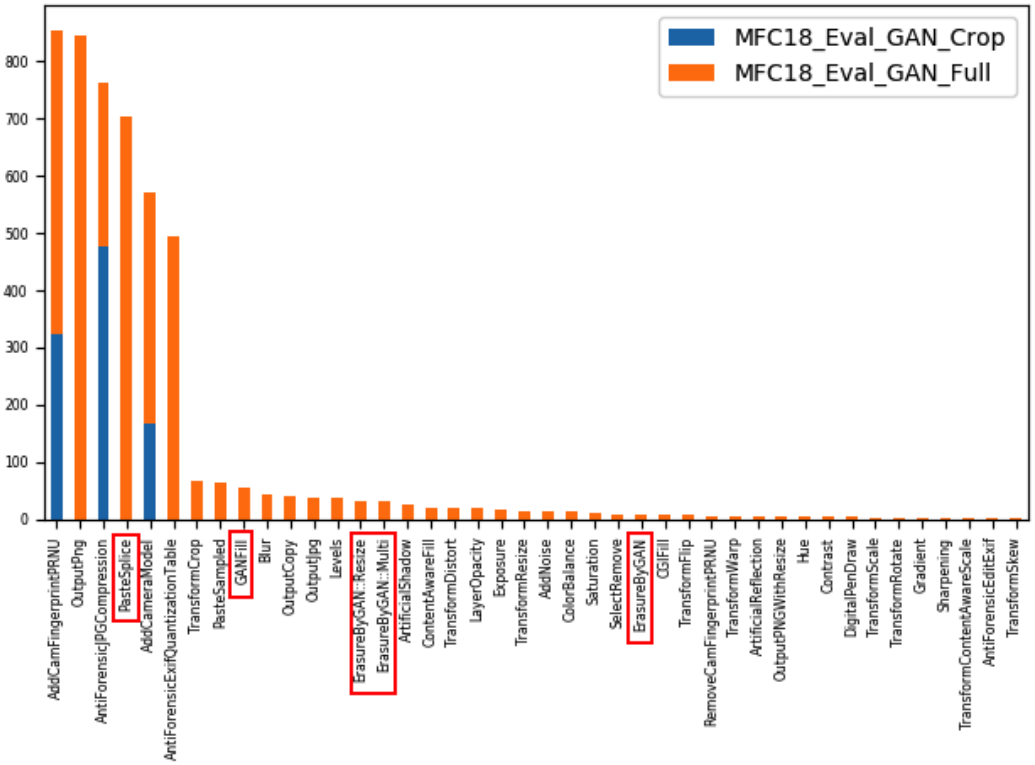


Figure 4.4. Manipulation type distribution on GAN image dataset; both GAN-based (marked in red) and non-GAN manipulation types were used for the target trial and unbalanced manipulation types was used in the MFC18 evaluation.

(a) Original image (b) Manipulation using both GAN & non-GAN

Figure 4.5. An example of the MFC18 GAN dataset (approved by IRB ITL-0018); (a) genuine image; (b) the manipulated image contains both GAN (e.g., ErasureByGAN) and non-GAN (e.g., ColorBalance) operations, so the authors suggest interpreting the GAN challenge results with caution.

5. Results

Although all MFC18 results are available with the different conditions, frameworks, and datasets described in Secs. 3 and 4, in this paper, the results for the *image-only* and the *video-only* conditions and the full scoring framework on the MFC18 *EvalPart1 dataset* are presented to describe results and demonstrate analysis methods.

5.1 Full Scoring Framework

A total of 25 teams from the academic and industrial sectors participated in the MFC18 evaluations, and the teams were allowed to have multiple submissions with their validated system outputs. Each team had the option to participate only in selected tasks and challenges. No team was in every task and challenge; hence the resulting summary in Table 5.1 has many missing values which resulted in analysis complications.

For performance measures, we used AUC (Area Under Curve) and CDR (Correction Detections at False Alarms = 0.05) for detection and verification tasks (MDL, SDL, EV, CV), while using MCC_o (Matthew Correlation Coefficient with an optimal threshold) for localization (MDL, SDL). Figures 5.1 and 5.2 illustrates ROC curves for detection and mask confusion matrix for localization, respectively.

Out of all the submissions, for each task and sub-task, we first identified the submissions under the evaluation conditions and frameworks for this paper. If there were multiple submissions for the team, we used the submission that had the highest system performance on the primary metric for the task or sub-task. For instance, in the case of team T12 , we identified a total of 27 submissions for the MDL-Image detection task. We then picked a submission with the highest AUC value for the detection. If two or more systems tied for the highest AUC value but had different CDR values, the system with the highest CDR value was chosen. For the MDL-Image localization, the T12 system with the highest MCC_o was

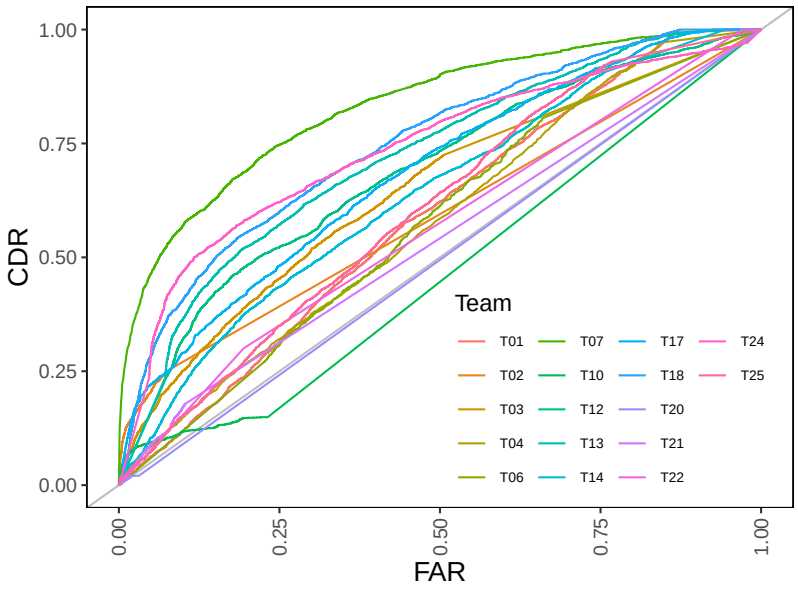


Figure 5.1. An example of ROC curves for participants in the MFC18 MDL-Image detection task; the x-axis is a false alarm rate (FAR) and the y-axis is a correct detection rate (CDR).

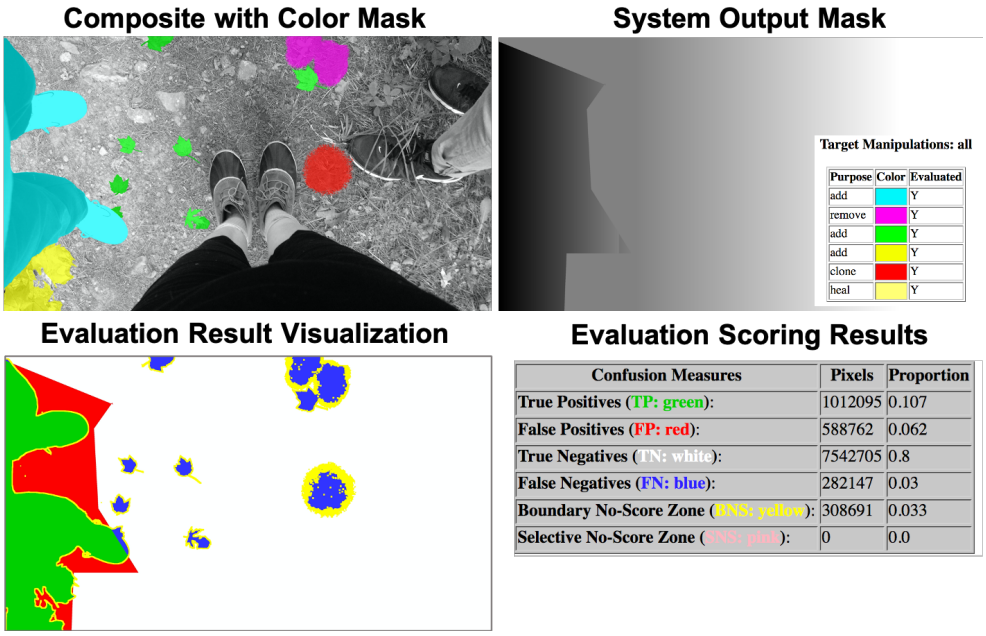


Figure 5.2. A graphical example of localization evaluations; top-left: ground-truth masks marked in color by different manipulation types, top-right: system output mask, bottom-left: Mask confusion matrix visualization, bottom-right: confusion matrix results for scoring localization.

chosen, which may be different than the T12 system chosen for the detection task; $pMCC_o$ indicates the probe localization metric while $dMCC_o$ is the donor localization metric. Ta-

ble 5.1 shows the top values for each team on the tasks; the primary metrics are described in Sec. 3. For all the metrics, a higher value is considered as better performance.

Table 5.1 shows team performance for the five evaluation tasks for both image and video detection and localization. For image evaluations, we had system submissions from 17 teams for MDL, 4 teams for SDL, 2 teams for EV and 3 teams for the PF and PGB tasks. For video evaluations, we had system submissions from 4 teams for the MDL task.

Table 5.1. Summary of MFC18 Image/Video Five Evaluation Task Results (Full Scoring Only)

Det: Detection, Loc: Localization, #Team: Number of participants

Team	MDL						SDL				EV		PF	PGB		
	Image			Video												
	Det		Loc	Det		Loc	Det		Loc							
	AUC	CDR	MCC_o	AUC	CDR	MCC	AUC	CDR	$pMCC_o$	$dMCC_o$	AUC	CDR	Re@300	sim_{NLO}	sim_{NO}	sim_{LO}
T01	0.59	0.06	0.21													
T02	0.59	0.21	0.05				0.68	0.31	0.30	0.31						
T03	0.65	0.16									0.85	0.48				
T04	0.58	0.08	0.07				0.77	0.62	0.36	0.33						
T05				0.59	0.13											
T06	0.57	0.06														
T07	0.83	0.47	0.15	0.59	0.12	0.09										
T09													0.90	0.57	0.79	0.29
T10	0.47	0.09	0.05													
T12	0.69	0.16	0.15								0.72	0.27				
T13	0.72	0.16														
T14	0.64	0.10	0.06				0.75	0.43	0.19	0.16						
T16													0.90	0.54	0.80	0.27
T17	0.69	0.22	0.05													
T18	0.75	0.28														
T19				0.51	0.01											
T20	0.50	0.04	0.00	0.40	0.03	-0.01										
T21	0.54	0.09	0.07													
T22	0.56	0.08	0.03													
T23							0.72	0.40	0.18	0.16			0.87	0.46	0.77	0.05
T24	0.74	0.29	0.26													
T25	0.61	0.07	0.10													
#Team	17	17	13	4	4	2	4	4	4	4	2	2	3	3	3	3

Table 5.2 illustrates performance on the two challenge tasks for both image and video. For image, 8 teams submitted their systems for the CV challenge with image training and testing sets and 9 teams for the GAN challenge. For video, 3 teams submitted their systems for the CV challenge with video training and testing sets and 5 teams for the GAN challenge.

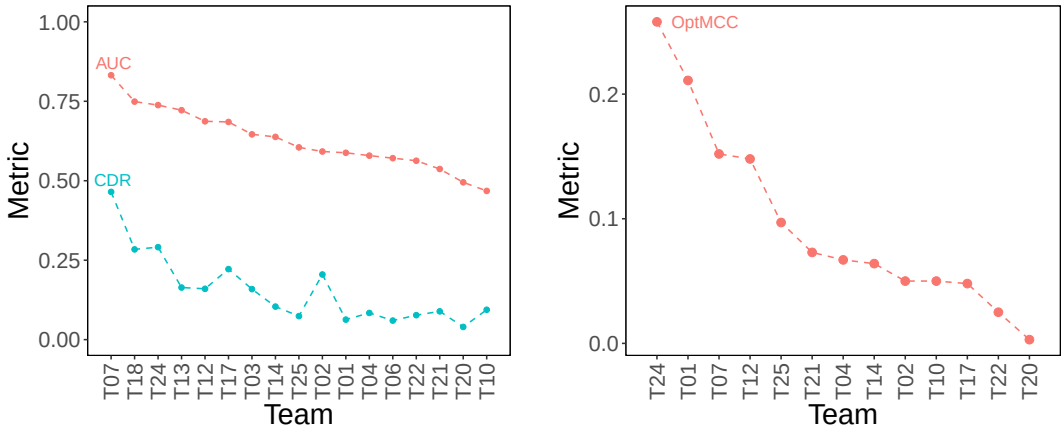
For The MDL-Image task, Figure 5.3-a ranks the 17 teams (ordered by AUC) for the MDL-Image detection task on 3,265 target trials and 14,156 non-target trials. The red line is the ranking for AUC, and the blue line is for the corresponding CDR. With obvious exceptions, the general trend of the two metrics performance is similar across the systems. Given the manipulation types, the results show that T07 has the highest AUC (0.83) followed by T18 (0.75) for the detection. Out of 17, as shown in Figure 5.3-b, only 13 teams participated in the localization evaluation, and the teams are ordered by optimum MCC values. For localization, T24 has the highest MCC_o (0.26) followed by T01 (0.21). For at least three teams, there are different systems selected for the detection and localization tasks than were selected for detection–which implies that a good detection system may not be necessarily robust to localize the manipulated region.

To demonstrate an analysis methodology, the block plot [20, 21] in Figure 5.4 was used to address three major questions: 1) how does the system perform on the different journal

Table 5.2. Summary of MFC18 Image/Video Challenge Results (Full Scoring Only)

Det: Detection, Loc: Localization, #Team: Number of participants

Team	CV				GAN				
	Image-Image		Video-Video		Image			Video	
					Det		Loc	Det	
	AUC	CDR	AUC	CDR	AUC	CDR	MCC_o	AUC	CDR
T01	0.77	0.55	0.55	0.18					
T04	0.51	0.07			0.61	0.03			
T07	0.65	0.32			0.69	0.10	0.07	0.68	0.12
T08					0.65	0.20			
T10					0.42	0.05	0.02		
T11								0.66	0.42
T12					0.66	0.12			
T15	0.87	0.77	0.60	0.36					
T17	0.55	0.07							
T18					0.49	0.09	0.01		
T19								0.68	0.24
T20	0.77	0.57	0.70	0.35					
T22					0.70	0.08	0.09	0.59	0.06
T24	0.81	0.57			0.79	0.59	0.37	0.74	0.58
T25	0.59	0.14			0.74	0.17			
#Team	8	8	3	3	9	9	5	5	5



(a) Detection (MDL-Image) (b) Localization (MDL-Image)

Figure 5.3. System performance ranking for MDL-Image; (a) detection performance ranking ordered by *AUC*; (b) localization performance ranking (OptMCC indicates MCC_o); the highest-performed system in detection is not necessary the highest-performed system in localization.

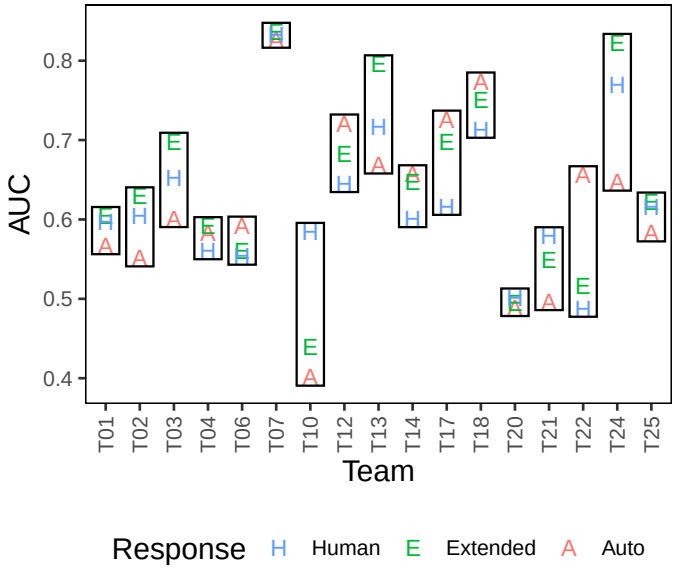
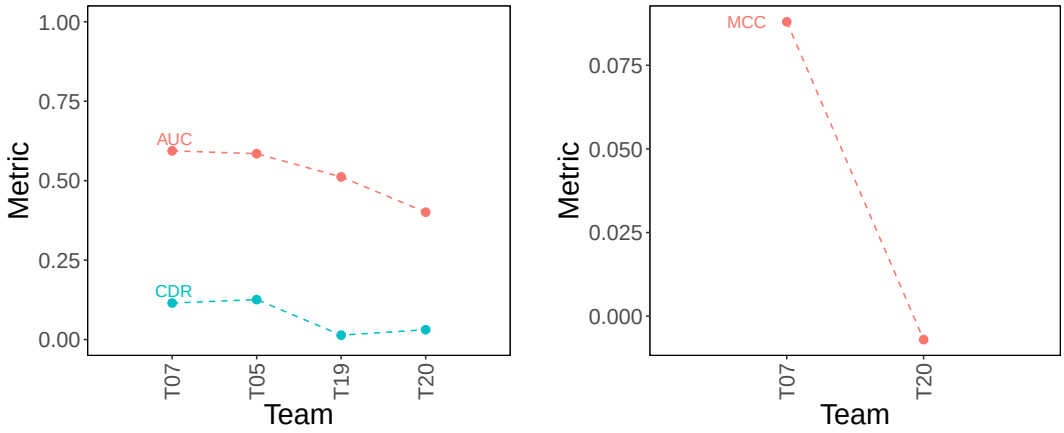


Figure 5.4. Effect of journal types across systems (MDL-Image case); The *x*-axis is the team and the *y*-axis is the mean *AUC*. The characters inside each bar represent the settings of the journal types. The journal type has a larger effect on some systems (e.g. T10, T18, T24) but is not as noticeable for other teams (e.g., T07 and T20).

types?, 2) is the conclusion robust, and 3) is the journal type an important factor?. The *x*-axis is the team, and the *y*-axis is the mean AUC based on the MDL. The characters inside each bar represent the settings of the journal types: Human-JT (H), Extended-JT (E), and Auto-JT (A). A taller bar indicates a larger impact of the journal type on that system compared to a smaller bar. The results showed that Human-JT and Extended-JT are easier to detect for some teams (e.g., T24) while Auto-JT are easier to detect for other teams (e.g., T18 and T22). The ranked list of the journal types is not consistent across the systems, and T07 has the highest AUC regardless the journal type. The journal type has a larger effect on some systems (e.g. T10, T18, T24) but is not as noticeable for other teams (e.g., T07 and T20).

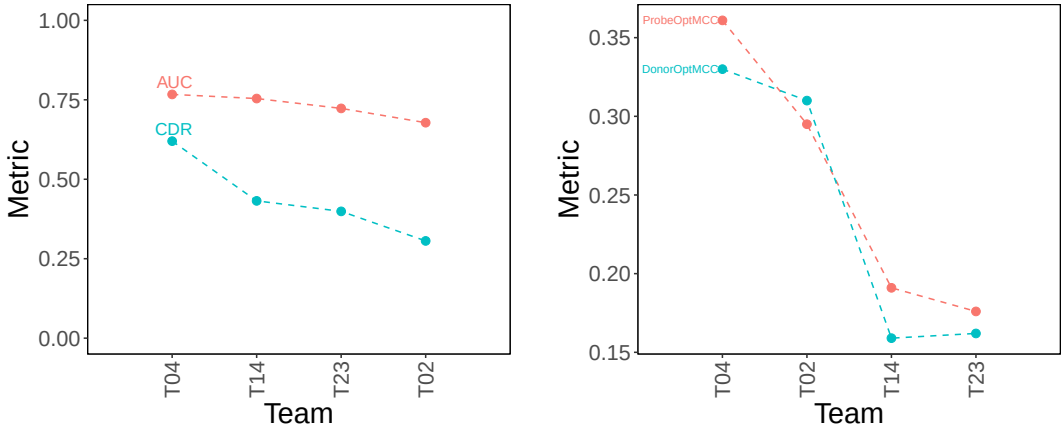
For the MDL-Video detection task, Figure 5.5-a ranks the results of the 4 systems (ordered by AUC) on 323 target trials and 713 non-target trials. The red line is the ranking for AUC, and the blue line is for the corresponding CDR. Over all manipulation types, T07 has the highest AUC (0.59) followed by T05 (0.59). Figure 5.5-b ranks the results of the 2 systems for the MDL-Video localization. Over all manipulation types, T07 has the highest MCC (0.09).

For the SDL task, we evaluated 4 systems on 1,327 target trials and 16,673 non-target trials. Figure 5.6 shows that T04 has the highest performance for both detection ($AUC = 0.77$) and localization ($pMCC_o = 0.36$, $dMCC_o = 0.33$) from probe and donor



(a) Detection (MDL-Video) (b) Localization (MDL-Video)

Figure 5.5. System performance ranking for MDL-Video; (a) detection performance ranking ordered by *AUC*; (b) localization performance ranking ordered by MCC_o; The team T07 has the highest AUC (0.59) and the highest MCC (0.09).

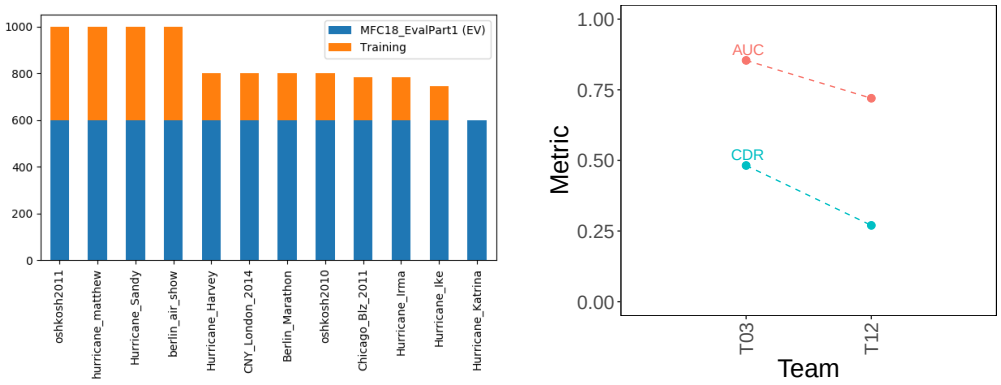


(a) Detection (SDL-Image) (b) Localization (SDL-Image)

Figure 5.6. System performance ranking for SDL; (a) detection performance ranking ordered by *AUC*; (b) localization performance ranking ordered by $pMCC_o$ (ProbeOptMCC and DonorOptMCC indicate $pMCC_o$ and $dMCC_o$, respectively.); The team T04 has the highest performance for both detection ($AUC = 0.77$) and localization ($pMCC_o = 0.36$, $dMCC_o = 0.33$).

masks, respectively.

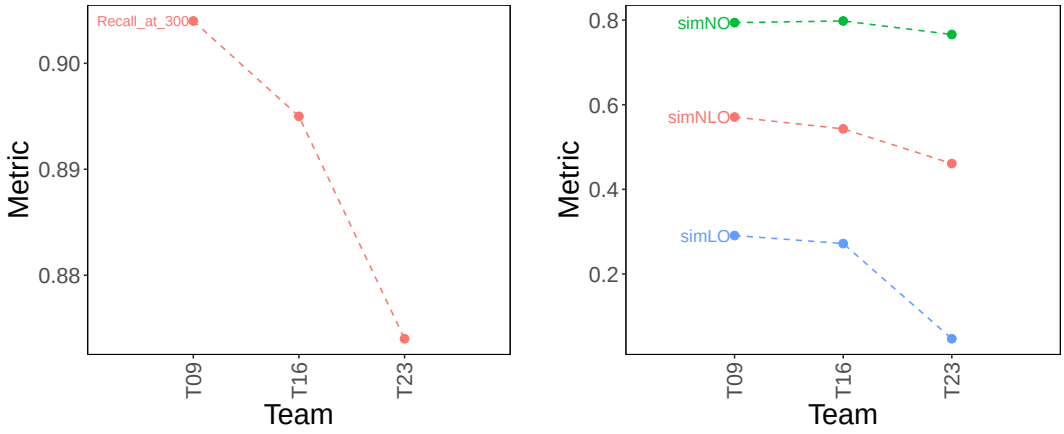
For the EV task, the MFC18 EvalPart1 dataset contained 12 different events. The training sets consisted of between 100 and 200 images while the testing sets contained 50 images for each event. The EV evaluation was conducted on 600 target pair-events and 6,600 non-target pair-events. The results in Figure 5.7 show that T03 has the highest AUC (0.85) for the EV task.



(a) EV data distribution (b) EV performance

Figure 5.7. The EV data and system performance ranking; (a) data distribution for MFC18 EvalPart1 and training data; (b) verification performance ranking ordered by *AUC*; The team T03 has the highest *AUC* value (0.85).

For the PF and PGB tasks, three teams participated in both the provenance tasks. The evaluations were conducted on 1,122 target probes and 1M world data. As illustrated in Figure 5.8, out of the three teams, T09 has the highest performance for both PF (Recall at 300: 0.90) and PGB ($sim_{NLO} = 0.57$).

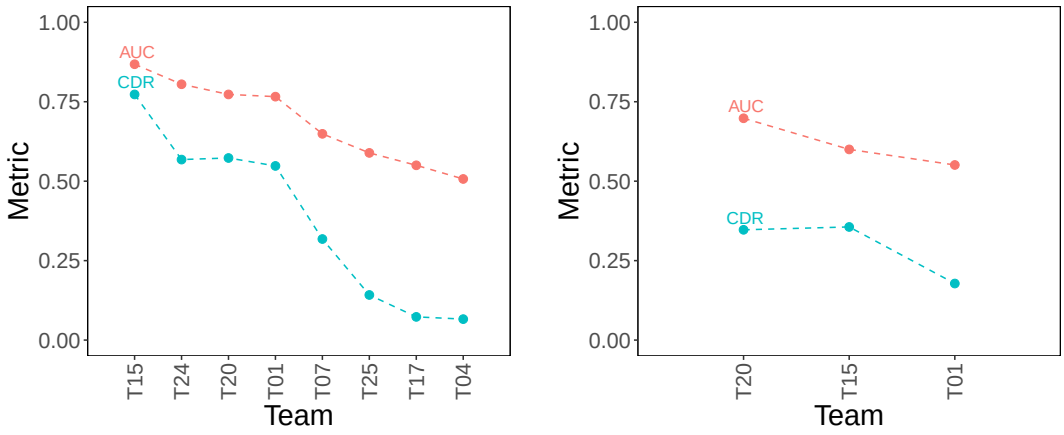


(a) PF (b) PGB

Figure 5.8. System performance ranking for provenance tasks; (a) PF performance ranking by *Recall*@300; (b) PGB performance ranking ordered by sim_{NLO}; The team T09 has the highest performance for both PF (*Recall*@300 = 0.90) and PGB ($sim_{NLO} = 0.57$).

In the CV challenge, we used 2,440 target and 2,835 non-target image trials from the condition where both training and testing sets are images (namely, Image-Image challenge). A total of 8 teams participated in the CV (Image-Image) challenge. As shown in Figure 5.9-a, the general trend of the two primary metrics is similar. The results show that T15 has

the highest AUC value (0.87) for the CV (Image-Image) challenge. For the CV (Video-Video) challenge, 3 teams participated and were evaluated their systems with 101 target 188 non-target video trials. Figure 5.9-b shows the general trends on the primary metrics. The results show that T20 has the highest AUC value (0.70) for the CV (Video-Video) challenge.



(a) CV (Image-Image) (b) CV (Video-Video)

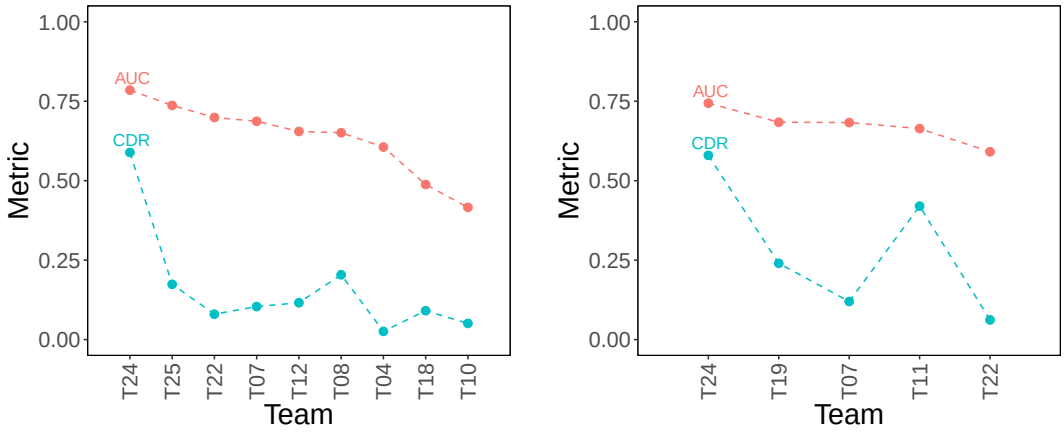
Figure 5.9. CV System performance ranking for CV; (a) Image-Image performance ranking ordered by *AUC*; (b) Video-Video performance ranking ordered by *AUC*; T15 has the highest ($AUC = 0.87$) for Image-Image while T20 has the highest ($AUC = 0.70$) for Video-Video.

Lastly, for the GAN challenge, Figure 5.10-a illustrates the evaluation results on the image dataset (855 target trials and 485 non-target trials) for detection while Figure 5.10-b shows the evaluation results on the video dataset (50 target trials and 68 non-target trials). As mentioned above, one of the drawbacks for the GAN dataset is that probes contain multiple manipulation types other than GAN-related operations which can mislead the interpretations of the GAN challenge results. Interpretation of the GAN challenge results should be done with caution.

5.2 Selective Scoring Framework (MDL-Image only)

For the MDL-Image task, there were at least 11 selective scoring queries under which teams could be scored. Teams indicated which scoring queries would be appropriate for each system. In this paper, for demonstration purposes, we selected 3 selective queries, namely, Clone, Remove, and Crop.

For the *Clone* selective scoring in the MDL-Image detection, Figure 5.11-a ranks the 6 teams (ordered by AUC from the full scoring). The red line is the ranking for AUC, and the blue line is for the corresponding CDR. With obvious exceptions, the general trend of the two metrics performance is similar across the systems. On the clone manipulation, the results show that T24 has the highest AUC (0.81) followed by T07 (0.80) for the detection. Out of 6 teams, 5 participated in the localization for the Clone selective scoring evaluation.



(a) GAN-Image detection (b) GAN-Video detection

Figure 5.10. System performance ranking for GAN; (a) GAN-Image performance ranking ordered by *AUC*; (b) GAN-Video performance ranking ordered by *AUC*; note that probes contain multiple manipulation types other than GAN-related operations which require caution to interpret the GAN challenge results.

For the localization, Figure 5.11-b shows that T24 has the highest MCC_o (0.27), followed by T07 (0.13).

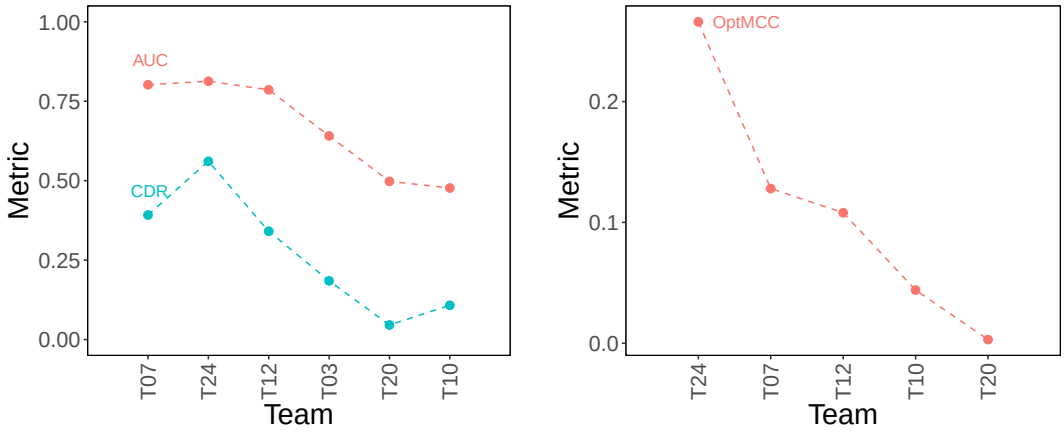
For the *Remove* selective scoring in the MDL-Image detection, Figure 5.12-a ranks the results of the 7 teams (ordered by AUC from the full scoring). On the Remove manipulation, the results show that T07 has the highest AUC (0.81) followed by T13 (0.79) for the detection.

For the *Crop* selective scoring in the MDL-Image detection, Figure 5.12-b ranks the results of the 6 teams (ordered by AUC from the full scoring). On the Crop manipulation, T07 has the highest AUC (0.83) followed by T22 (0.82) for the detection.

5.3 NC17 and MFC18 Comparison

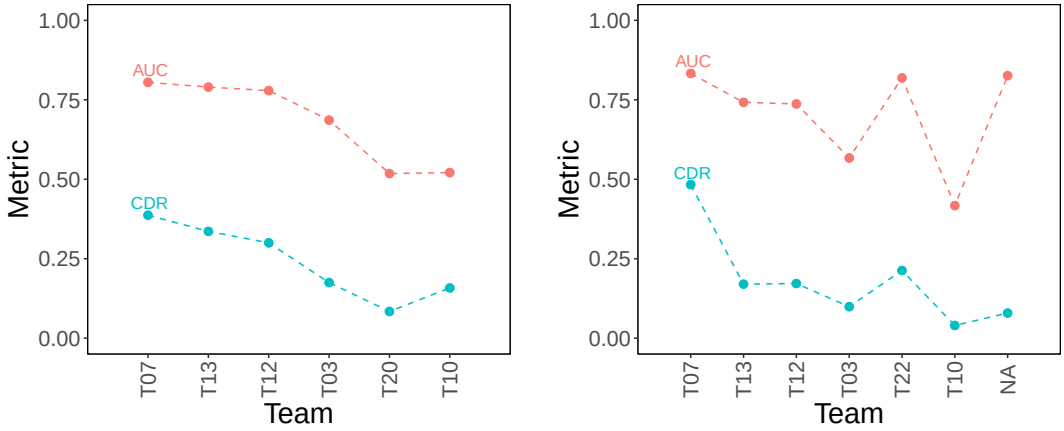
Prior to the MFC18 evaluations, we introduced the 2017 Nimble Challenge (NC17) to advance the state-of-the-art for media forensics technologies that automatically determine the region and type of manipulations in imagery [22]. In this section, we briefly discuss the comparison of results between the NC17 and MFC18 evaluations. For the performance comparison, we picked the highest performance value for each task as illustrated in Figure 5.13. Note that different datasets and different system submissions between the NC17 and MFC18 evaluations were used in this comparison. The paper [18] introduced both NC17 and MFC18 datasets in detail as well as how they were built to support the research community.

The results in Figure 5.13 indicate that the MDL and SDL system performance in the MFC18 evaluation was slightly improved while the PF and PGB performance in the MFC18 had a significant improvement compared to the prior NC17 evaluation. For the Prove-



(a) Detection (MDL-Image: Clone) (b) Localization (MDL-Image: Clone)

Figure 5.11. System performance ranking for selective scoring (MDL-Image: Clone); (a) detection performance ranking ordered by *AUC*; (b) localization performance ranking; T24 has the highest performance for both detection ($AUC = 0.81$) and localization ($MCC_o = 0.27$).



(a) Detection (MDL-Image: Remove) (b) Detection (MDL-Image: Crop)

Figure 5.12. System performance ranking for selective scoring (Remove and Crop); (a) Remove detection performance ranking ordered by *AUC*; (b) Crop detection performance ranking ordered by *AUC*; T07 has the highest performance for both the Remove and Crop detection ($AUC = 0.81$ and $AUC = 0.83$, respectively).

nance Filtering (PF) comparison, the metric *Recall*@200 was used for the comparison since *Recall*@300 was not available in the NC17 evaluation; hence the PF highest values between Table 5.1 and Figure 5.13 are not the same.

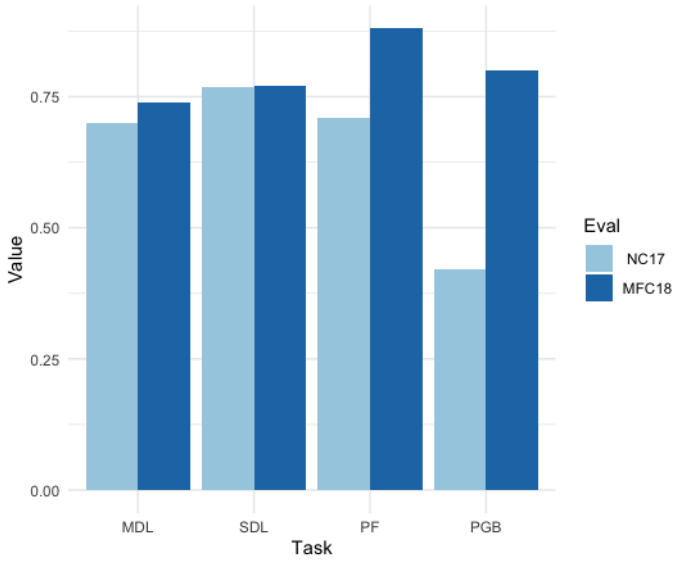


Figure 5.13. Tasks performance comparison of the NC17 and MFC18 evaluations (different systems and different datasets were used for the comparison).

6. Conclusions

In this paper, we presented an evaluation framework, analysis methodology, and how it was applied to the MFC18 results with a full scoring and selective scoring evaluation. For MFC18, we defined the five tasks (MDL, SDL, EV, PF, and PGB) and two challenges (CV and GAN) with associated performance metrics. A total of 25 teams participated in the MFC18 evaluations for the image and video tasks, and the experiments were conducted for comparative analysis using a structured evaluation framework to assess the accuracy and robustness of a system. The MFC18 evaluation produced a large number of diverse datasets (both manually and automatically) for evaluating the defined tasks. We also provided a ranked list of system performance for each task and demonstrated the factor effect using the journal type. Our results showed that the highest AUC value is 0.83 (T07) for the MDL-Image detection, and MCC_o is 0.26 (T24) for the MDL-Image localization. For video, we found the highest AUC value is 0.59 (T05 and T07) for the MDL-Video detection. We found a general difference between detection and localization performance and that some teams had different systems with best performance for each task, implying that a good detection system may not be necessarily robust to localize the manipulated region. The journal type has effect on some systems.

We hope that the MFC18 evaluation provides researchers insight on their system performance and direction of their system development. Further, it is our hope that it provides the computer vision community the data and evaluation foundation needed for advancing media forensics technology.

Acknowledgments

Authors would like to acknowledge DARPA (Defense Advanced Research Projects Agency) who funded this program as well as PAR Government Systems, RankOne, Rochester institute of Technology, Drexel University, and the University of Michigan for their data collection efforts. We thank David Joy and August Pereira (former employees of NIST/ITL) for their contributions to the scoring code and server. The authors would also like to thank Dr. John Henry Scott for his invaluable comments and suggestions. NIST conducted this work under NIST Interagency Agreement Number 1505-774-08-000 with Institutional Review Board (IRB) approval (ITL-0018).

References

[1] NIST MediFor Team Media forensics challenge 2018 evaluation plan. [Accessed 21-May-2020] Available at https://www.nist.gov/system/files/documents/2018/10/30/mfc2018evaluationplan-clean3_werb.pdf.

[2] Stamm MC, Wu M, Liu KR (2013) Information forensics: An overview of the first decade. *IEEE Access* 1:167–200. https://doi.org/10.1109/ACCESS.2013.2260814. Available at https://doi.org/10.1109/ACCESS.2013.2260814

[3] Scherhag U, Rathgeb C, Merkle J, Breithaupt R, Busch C (2019) Face recognition systems under morphing attacks: A survey. *IEEE Access* 7:23012–23026.

[4] Bik EM, Fang FC, Kullas AL, Davis RJ, Casadevall A (2018) Analysis and correction of inappropriate image duplication: the molecular and cellular biology experience. *Molecular and Cellular Biology* 38(20). https://doi.org/10.1128/MCB.00309-18. https://mcb.asm.org/content/38/20/e00309-18.full.pdf Available at https://mcb.asm.org/content/38/20/e00309-18

[5] Christlein V, Riess C, Jordan J, Riess C, Angelopoulou E (2012) An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on Information Forensics and Security* 7(6):1841–1854.

[6] Meena KB, Tyagi V (2019) *Image Forgery Detection: Survey and Future Directions* (Springer Singapore), , , , pp 163–194. https://doi.org/10.1007/978-981-13-6351-1_14. Available at https://doi.org/10.1007/978-981-13-6351-1_14

[7] Columbia image splicing detection evaluation dataset. [Online; accessed 30-June-2020] Available at http://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm.

[8] Thakur T, Singh K, Yadav A (2018) Blind approach for digital image forgery detection. *International Journal of Computer Applications* 179(10):34–42. https://doi.org/10.5120/ijca2018916108. Available at http://www.ijcaonline.org/archives/volume179/number10/28839-2018916108

[9] Verdoliva L (2020) Media forensics and deepfakes: an overview. 2001.06564.

[10] Yang P, Baracchi D, Ni R, Zhao Y, Argenti F, Piva A (2020) A survey of deep

learning-based source image forensics. *Journal of Imaging* 6(3). https://doi.org/10.3390/jimaging6030009. Available at https://www.mdpi.com/2313-433X/6/3/9

[11] de Oliveira AA, Ferrara P, De Rosa A, Piva A, Barni M, Goldenstein S, Dias Z, Rocha A (2016) Multiple parenting phylogeny relationships in digital images. *IEEE Transactions on Information Forensics and Security* 11(2):328–343.

[12] Moreira D, Bharati A, Brogan J, Pinto A, Parowski M, Bowyer KW, Flynn PJ, Rocha A, Scheirer WJ (2018) Image provenance analysis at scale. *IEEE Transactions on Image Processing* 27(12):6109–6123.

[13] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems 27*, eds Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (Curran Associates, Inc.), , pp 2672–2680. Available at http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[14] Fawcett T (2006) An introduction to roc analysis. *Pattern Recognition Letters* 27(8):861 – 874. https://doi.org/https://doi.org/10.1016/j.patrec.2005.10.010. ROC Analysis in Pattern Recognition Available at http://www.sciencedirect.com/science/article/pii/S016786550500303X

[15] Wikipedia contributors (2020) Matthews correlation coefficient — Wikipedia, the free encyclopedia, https://en.wikipedia.org/w/index.php?title=Matthews_correlation_coefficient&oldid=958781090. [Online; accessed 30-June-2020].

[16] Papadimitriou P, Dasdan A, Garcia-Molina H (2010) Web graph similarity for anomaly detection. *Journal of Internet Services and Applications* 1(1):19–30. https://doi.org/10.1007/s13174-010-0003-x. Available at https://doi.org/10.1007/s13174-010-0003-x

[17] Robertson E, Guan H, Kozak M, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) Manipulation data collection and annotation tool for media forensics. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, , pp 29–37.

[18] Guan H, Kozak M, Robertson E, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, , pp 63–72. https://doi.org/10.1109/WACVW.2019.00018. Available at https://doi.org/10.1109/WACVW.2019.00018

[19] ERobertson Userguide for maskgen journaling tool. [Accessed 21-May-2020] Available at https://gitbub.com/rwgdrummer/maskgen/blob/master/doc/MediForJournalingTool-public.pdf.

[20] J J Filliben and A Hecket Engineering statistics handbook (5.5.9. an eda approach to experimental design). [Accessed 21-September-2020] Available at http://www.itl.nist.gov/div898/handbook/pri/section5/pri59.htm.

[21] JJFilliben (1981) Dataplot—an interactive high-level language for graphics, nonlinear fitting, data analysis, and mathematics. *ACM SIGGRAPH Computer Graphics* 15(3). https://doi.org/10.1145/965161.806807. Available at https://dl.acm.org/doi/

abs/10.1145/965161.806807#sec-ref
[22] NIST MediFor Team Nimble challenge 2017 evaluation plan. [Accessed 21-May-2020] Available at https://www.nist.gov/system/files/documents/2017/09/07/nc2017evaluationplan_20170804.pdf.

Appendix A: Change Log

Revision 1 Release – February 19, 2021

- Made editorial changes throughout the report.
- Revised sentences in Section 1 with clarification.
- Fixed spelling mistakes in Reference.
- Revised Acknowledgments